DOCUMENT RESUME

ED 064 347 TM 001 514

AUTHOR Mandeville, Garrett K.
TITLE A Comparison of Three Methods of Analyzing Dichotomous Data in a Randomized Block Design.
NOTE 38p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Comparative Analysis; Data Analysis; *Mathematical Models; Models; Research; *Statistical Analysis; *Tables (Data); *Test Reliability

ABSTRACT

Results of a comparative study of F and Q tests, in a randomized block design with one replication per cell, are presented. In addition to these two procedures, a multivariate test was also considered. The model and test statistics, data generation and parameter selection, results, summary and conclusions are presented. Ten tables contain the study data. (DB)

ED 064347

A Comparison of Three Methods of Analyzing Dichotomous Data in a Randomized Block Design

Garrett K. Mandeville, University of South Carolina

Introduction and Review of Literature

Many situations arise in behavioral research where the data fit nicely into the randomized block paradigm. For example, when information is available on an antecedent variable such as aptitude, subjects may be "blocked" on the basis of these scores to gain precision in comparing effects of the independent variable, e.g., teaching methods, on the dependent variable, e.g., achievement. Often this randomized block design would be an improvement over the completely randomized design. When a single subject responds to a series of trials or to varying treatment conditions given in random order the term repeated measures design is commonly used in the behavioral literature. In each case the experimenter wishes to compare the strength of response for the treatments, trials, etc., or test or estimate some contrast of their effects. The methods of analysis are the same, however, and this is why the terminology randomized block paradigm was used above.

When the response variable is continuous, analysis of variance (ANOVA) or a multivariate procedure would probably be used to analyze the data. A recent study by Porter and McSweeney (1970) has considered the advantages of blocking in situations where a non-parametric analysis is to be performed. It is this writer's contention, however, that many situations occur in behavioral research where the measurement is of such quality as to invalidate any of these techniques. Of particular interest in this study are situations where the dependent variable is dichotomous. A few examples are learning trails in which the subject makes either the correct or incorrect associations, maze running in which a rat turns right or left, problem solving where the problem is solved or not solved and attitude surveys in which the respondent agrees or disagrees. Sometimes the crude dichotomous response is obtained to facilitate speed. Numerous examples of dichotomous dependent variables in psychological experiments are

TM 001 514

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

given in Seegar and Gabrielson (1968, p. 270). The question of how to analyze data in a randomized block design when the dependent variable is dichotomous is not discussed with any frequency in the applied statistical literature.

Cochran (1950) presented the Q test as a procedure for testing the hypothesis that in each block, the response probabilities are constant for the various treatment levels. (Note that the terms blocks and treatments will be used for convenience here; the reader should understand that blocks may represent a set of subjects grouped together or multiple measurements on the same subject, etc., and that treatments, which might be more appropriately called treatment levels, are the conditions for which comparisons of effects are desired. Also, for definiteness, let us assume that there are I blocks and J treatments.) The Q statistic is based on a randomization argument, i.e., if the treatments are no different, the u_i positive responses in block i could have, with equal probability been arranged in any of the $\binom{J}{u_i}$ ways. Under randomization, the responses have the same variances and because of symmetry the covariances between any response pairs are also equal. The sum of squared deviations among the J treatment totals is, for large samples, shown to approximate a multiple of the χ^2 distribution.

Let us clarify the hypothesis tested with Q. Let ϕ_{ij} represent the "success" probability associated with the application of the jth treatment in the ith block. Then the null hypothesis of the Q test, is $Ho_1 : \phi_{i1} = \phi_{i2} = \ldots = \phi_{iJ}$ for i=1, 2,....,I. Note that the hypothesis of ANOVA is of lesser magnitude, i.e., $Ho_2: \bar{\phi}_{.1} = \bar{\phi}_{.2} = \ldots = \bar{\phi}_{.J}$ where $\bar{\phi}_{.j}$ is the average success probability for the jth treatment. It should be clear that Ho_1 is a composite of Ho_2 and the hypothesis of no treatment x block interaction. Seegar and Gabrielson (1968) consider data sets with varying amounts of treatment x block interaction and observe that, as one would expect, the Q test does have power to detect situations where interaction exist, but Ho_2 is true. It is probably true, partly because of the unfortunate way Q is handled in secondary sources, that resear-

chers do not realize the subtle differences in these hypotheses. Tukey's test for non-additivity would provide one method for detecting an interaction of the multiplicative type, but would probably be of limited usefulness in most applications in the behavioral sciences where blocks are random. Other suggestions are given in Draper and Porter (1970).

When Cochran introduced the Q test, he presented some results comparing probabilities obtained using the χ^2 distribution with the exact probabilities obtained using randomization. The problems he studied were kept small to facilitate computation, the number of treatments ranging from 3 to 5 with from 6 to 16 blocks. In addition to computing Q for these data sets, the F ratio for a randomized blocks ANOVA was also computed. Cochran also computed Q' and F' which were values of these statistics corrected for continuity and they were obtained by finding the next smaller value of the statistic and averaging the two. For each of these statistics, then the appropriate table was entered and the probability of a larger value was obtained. These were then converted to a percentage error by computing

$$100\ (\text{Tabular P}-\text{True P})/(\text{True P})$$

Cochran stated that the F' was decidedly better than F and so he did not present results for F. Of the other three statistics, Q was suggested to be preferred. It exhibited a negative bias, i.e., a tendency to underestimate the true probability when true P was in the range of .2 to .02 and, overestimate the true probability (corresponding to a conservative test) for true probabilities below .02. The corrected χ^2 had a positive bias over the whole range of true probabilities while F' had a tendency towards a negative bias.

Although it may strike the reader as being strange that Cochran would even consider the F test in this situation, the following quote is illuminating: "I had once or twice suggested to research workers that the F test might serve as an approximation

even when the table consists of 1's and 0's'...this suggestion was received with incredulity, the objection being made that the F test requires normality, and that a mixture of 1's and 0's could not by an stretch of the imagination be regarded as normally distributed. The same workers raised no objections to a χ^2 test, not having realized that both tests require, to some extent, an assumption of normality, and that it is not obvious whether F or χ^2 is more sensitive to this assumption" (1950, p.262). Cochran further justifies consideration of F because of the widespread interest in the application of analysis of variance to non-normal data and, although his results are discouraging with regard to the F test, he notes that the application of F test to more complex tables should be kept in mind.

One important point which Cochran makes, which has been overlooked by most secondary sources is that Q is invariant under deletion of what will be termed here as "non-informative" blocks, i.e., blocks with responses of all 0's or 1's. Therefore, the term "large samples" must be used with caution and many statements in textbooks such as McNemer (1962) and Siegel (1956) are misleading in this regard.

A more recent study of the sampling distribution of Q for small samples was done by Tate & Brown (1964, 1970). These writers clarify the fact that Q is not changed upon deletion of non-informative blocks but point out that the F test is changed by this deletion. This is because the degrees of freedom of the residual mean square will change with the number of blocks in the experiment thus affecting both its value and the reference F distribution. Using data from Fleiss (1965), Tate and Brown compare results using F and Q and the exact probabilities from randomization. When no rows are deleted, the percent errors of F and Q are similar for the full table of the Fleiss data but when non-informative rows are deleted, this one set of data suggests that the F test has a negative bias. Tate and Brown go on and tabulate the exact distribution of Q for designs with 3 treatment and from 3 to 12 blocks; 4 treatments and from 2 to 8 blocks; 5 and 6 treatments and from 2 to 5 blocks. The tables are

presented in Tate and Brown (1964). By comparing Q to the exact probabilities for these tables these writers observe a negative bias in Q, but suggest that, when the product of the number of (informative) blocks and treatments is 24 or more, the approximation using Q is probably sufficient for most practical work. This is based on the observation that, for the distributions they considered for which this was true, median percent errors were in the range of 12% to 20%. For situations in which the product is less than 24, the exact tables of Tate and Brown may be used. Tate and Brown give no comparison of F and Q for these distributions.

The other rather extensive study reported in the literature concerning Q is the one already mentioned by Seegar and Gabrielson. These researchers were interested in extending the Q test to a situation in which there was more than one measurement available for each treatment-block combination. An extension of Q to this situation was presented and it was compared to the ANOVA F test using the treatment x block mean square as the error term. Although they did not consider it, a point in favor of the F test in this design is that it allows both treatment effects and treatment x block interactions to be tested whereas the extended version of the Q test which they suggested tested the same hypothesis as Cochran's original test. The results of this study suggest that where Ho_1 is true, Q (or the modification) can be used when the product of the number of blocks and the number of replications per treatment is in the range 10 - 20. The above test, however, requires that treatment x block interactions be null for otherwise Ho_1 is false and the Q test will have power to detect this. In these situations the F test, to test Ho_2, will still provide a reliable answer for from 5 - 10 subjects. Seegar and Gabrielson point out that when Ho_1 (and therefore Ho_2) is true, the F test is as good as Q. The arcsin transformation was found to provide results which were very similar to those of the F test.

Although Cochran's results with F (and F') suggest that it leads to underestimates of the true probabilities, it needs to be said that Seegar and Gabrielson did not find

this to be true in their study. There is a subtle difference in the methods used by Seegar and Gabrielson and by this writer on the one hand and by Cochran and Tate and Brown on the other. The method used here was to define critical regions using the reference F or χ^2 distributions and tallying the instances of a significant value for the statitstic. Thus, what becomes important is how the empirical (discrete) and theoretical (F or χ^2) cumulative distributions compare at the selected points. In the study to be reported here, the 90th, 95th, 97.5th and 99th percentiles of the F and χ^2 distributions were used to designate lower boundaries of critical regions for α = .10, .05, .025 and .01. These were selected because they cover common significance levels employed by educational researchers.

This writer decided to take a closer look at the properties of the F test in a randomized block design with one replication per cell. This comparative study of the F and Q tests seemed in order since the only comparisons available in the studies cited above were limited to the few examples of Cochran and Tate and Brown and the few simulations for one replication of Seegar and Gabrielson. The research is restricted to one replicate because it was anticipated that researchers with more than one observation per cell would use an F test to allow for a test of interaction.

Some writers have suggested that recent results of Hsu and Feldt (1970) and Lunney (1970) for ANOVA with dichotomous data in independent cells designs can be applied to ANOVA in designs where the data are correlated. Because of the differences in fixed and mixed ANOVA models, this writer suggests that a certain amount of caution be exercised here. It is the feeling of this writer that the F test should be the recommended procedure if it simply provides results which are as good as Q because; (1) it is more flexible than Q and, therefore, more promising for complex designs, (2) it is so frequently used, and therefore probably fairly well understood by educational researchers and (3) many computer programs are available to facilitate the

analysis. For researchers who are mainly concerned with differences in average success proportions, some form of F test should be recommended over Q because of the power of the Q test if treatment x block interactions exist.

The Model and Test Statistics

A common specification of the model is $Y_{ij} = \mu + a_i + \beta_j + \varepsilon_{ij}$ where the a_i are block effects and are randomly sampled from a normal population with mean zero and variance σ_A^2, the β_j represent fixed treatment effects ($\Sigma_j \beta_j = 0$) and residual variation is included in the ε_{ij}. Tne ε_{ij} are from a normal population with mean zero and variance σ_ε^2 and the a_i and the ε_{ij} are independent (that is, within and between sets). This model leads to correlations between measurements within a block which are the same and this, therefore, becomes an assumption in the "univariate" ANOVA solution of the problem.

Without going into the detail of reproducing computational formulas, which are well known to most, some comparisons will be drawn between the Q and F statistics. In order to compute the Q and the F statistic, the following two sums of squares were obtained: SS_T, the sum of squares due to treatments, and SS_E, the error sum of squares Then formulas for Q and F are $Q = I(J-1)\ SS_T/[SS_T + SS_E]$ and $F = \frac{MS_T}{MS_E} = [SS_T/(J-1)]/[SS_E/(I-1)(J-1)]$. The null distribution of the Q static approaches χ^2_{J-1} as I increases (this assumes that some "informative" vectors have a positive probability) and the F statistic was compared to the F distribution with J-1 and (I-1)(J-1) degrees of freedom. Manipulations carried out to allow comparison of F and Q yield

$$Q/(J-1) = \frac{MS_T}{[(I-1)\ MS_E + MS_T]/I} .$$

This statistic, of course, can be referred to the $F_{J-1,\infty}$ distribution. Paralleling a discussion of D'Agostino (1971) we observe that the denominators of F and Q/(J-1) differ only slightly, and we will expect that, for large numbers of blocks, test size results for the two methods will be very similar. However, it appears as though, when treatment affects are non-null, the denominator will be inflated and power of Q relative to F, will be lowered.

A more general model than the one above is to consider the observations in a block, the $Y_{i1}, Y_{i2}, \ldots Y_{iJ}$, to be a vector observation from a J-variate normal

distribution with mean vector $\underset{\sim}{\mu}_y$ and general covariance matrix $\ddagger_y$. The hypothesis of interest, in this more general case, is that the elements of $\underset{\sim}{\mu}_y$ are equal, i.e., that $\mu_{y1} = \mu_{y2} = \ldots = \mu_{yJ}$. The problem is solved by transforming the Y's to a space of J-1 contrasts (differences are the most straightforward) using a transformation matrix $\underset{\sim}{C}$ which has J-1 independent contrasts as rows. Then $\underset{\sim}{X} = \underset{\sim}{C}\underset{\sim}{Y}$ and the hypothesis above becomes $\underset{\sim}{\mu}_x = \underset{\sim}{0}$ which can be tested using Hotelling's T^2 statistic, $T^2 = I\,\overline{X}'\,S_x^{-1}\overline{X}$. Actually a simpler computational form is given by Rao (1965) as $T^2 = I(I-1)\,[|\,\underset{\sim}{A}_x + \underset{\sim}{\overline{X}}\,\underset{\sim}{\overline{X}}'\,|\,/\,|\,A_x\,| - 1]$ where $\underset{\sim}{A}_x$ is the sum of products matrix for the transformed variables. The multivariate statistic used here was $[I-J+1]\,T^2/[(I-1)(J-1)]$ which is distributed as an F with J-1 and I-J+1 degrees of freedom if the hypothesis is true. The reader is reminded that I in the above formulas is the number of blocks not the identity matrix.

Data Generation and Parameter Selection

The generation of a dichotomous response vector requires a researcher to specify a model which allows various degrees of dependency to be manifest in the responses. If this study were dealing with continuous variables, the multivariate normal distribution would probably be the model used because of the many statistical procedures which assume that the data are sampled from it. In this study a multivariate normal distribution was assumed to be latent in the data, i.e., the assumption was that underlying each response was a normally distributed continuous random variable. This continuous variable was then compared to a fixed "cutting score" and if the response surpassed it a one was recorded; otherwise the response was taken to be a zero. The cutting score was determinded by the success proportion associated with the particular measurement involved.

The problem of selecting the multivariate normal distributions to be used for the generation of the latent variables amounted to selecting correlation matrices, because the means and variances for the dichotomous variables are determined by the success proportions used. When only two measurements are made in each block, only one correlation needs to be specified and the four values taken for this parameter were .0, .2, .5 and .8. The use of the zero correlations, or no association between the two measurements, will provide information on the extent to which applying an analytic technique for correlated data will penalize the researcher when in fact the data are unrelated. The use of a maximum correlation of .8 was justified because it is unusual when larger correlations occur in practice in educational research.

For more than two measurements, however, more than one measure of association needs to be selected. For three measurements, for example, there are three pairwise correlations which need to be specified. For the major portion of the study, these pairwise correlations were all taken to be equal and again the values of .0,

.2, .5, and .8 were used. What this means is that for the latent variables for most of the results presented here, the ANOVA assumption of equal correlations among the variables was satisfied. For the portion of the investigation dealing with type I error, i.e., when the success proportions for the treatments were equal, the population covariance matrices for the dichotomous variables also satisfied this assumption. However, when power was investigated, i.e., the success proportions were not the same for all treatments, the covariance matrix does not satisfy the pattern assumption of equal variances and covariances.

The number of treatments in a block was taken to be 2, 3, 6 and 10. Again, it was thought that this would provide a range for this parameter which would include most practical cases in educational experimentation. An exception would be a situation where each response was an item response in a test. In this case, of course, tests with more than ten items would be commonplace. However, it is not generally of interest to a researcher to determine whether the items in a test are of the same difficulty. The number of blocks was taken to be 5, 10, 20, and 30. It was anticipated that whatever large sample effects which were to be observed would be manifest for samples of size 30.

For the investigation of type I error, the null hypothesis is true, and the success proportions for all treatments are the same. To span the range of success proportions from 0 to 1 is unnecessary since by redefining success and failure, results for the range 0 to .5 can be applied to the range .5 to 1. The values of , the constant success proportion, which were used in this study were .1, .3 and .5. For success proportions smaller than .1, it is anticipated that, unless sample sizes are extremely large, a researcher should probably consider some alternative method of analysis.

Taking all combinations of the four values of J, the number of treatments (2, 3, 6, and 10) and I, the number of blocks (5, 10, 20, and 30) sixteen different designs

were initially to be considered. Due to limitations on computer time, the 10 treatment by 30 block design was eliminated. For each of the remaining fifteen designs, simulations were run for each level of ρ, the correlation among the measurements in a block (.0, .2, .5 and .8) in combination with each value of ϕ, the constant success proportion (.1, .3, and .5). Therefore, twelve simulations occurred for each design. In addition, for 3 and 6 treatments, a non-patterned correlation matrix was constructed and this was used in conjunction with each of the three ϕ values. These runs were limited to designs with 10 and 20 blocks.

To describe how the simulation took place the following listing of the steps involved in the determination of empirical test size is presented:

1. Values of I, the number of blocks, and J, the number of treatments, were set.
2. A correlation matrix R either constant with intercorrelations of .0, .2, .5, and .8 or a non-patterned matrix in a few special cases, was specified.
3. The vector of success proportions with all elements equal to ϕ = .1, .3 or .5 was selected.
4. A sample vector was generated from a multivariate normal distribution with covariance matrix R.
5. Each response was converted to a one if it surpassed the $(1-\phi)$th percentile of the standard normal distribution. Thus for ϕ=.3, any latent variables larger than the 70th percentile of the standard normal distribution, i.e., larger than .52, were converted to a one; otherwise a zero was recorded. In this way, each continuous response vector was converted to a vector of 0's and 1's.
6. The quantities necessary for computation of the three test statistics were accumulated.
7. Steps 4 - 6 were applied for data for each of the I blocks.

8. The values of the three statistics, Q, F, and M were obtained.
9. The boundaries for frequency distributions for each of the three statistics were computed. Of importance here are the 90th, 95th, 97.5th and 99th percentiles which were obtained for each of the three reference distributions. This step in the program was bypassed after the first data set was generated.
10. The computed statistics were cast into the frequency distributions set up in step 9.
11. Steps 4 - 10 were performed 1000 times.

The resulting empirical proportions above each of these percentiles are to be taken as estimates of true type I error (test size) for the test procedure.

For consideration of power the only alteration in the procedure was that, in step 3, a non-null proportion vector was read into the computer which would, under certain normal theory considerations, yield power of .60 and .80 for type I error of .05 or .01. Thus, the four combinations of α and $1-\beta$ required four simulations for a given design and R matrix.

Before discussing the results a word about how non-informative blocks were handled in this investigation is in order. From a logical point of view, these data vectors support the null hypothesis since, within them, no differences between the treatments are manifest. If the measurement scale had not been so crude, these scores would probably have been "closer together" than scores for blocks which exhibited variation. The retention of these vectors, which causes the estimated probabilities to increase and, therefore, the statistical tests to become more conservative, is one way to take account of such data. The retention of degrees of freedom in the denominator of F for these vectors is along the lines of Cochran's suggestion that, since F as he computed it tended to be liberal and χ^2, with essentially infinite degrees of freedom in the denominator, tended to be conservative, possibly some artificial number of these "non-informative" vectors would yield an empirical distribution which

provided a better fit to its theoretical counterpart. This researcher was not contriving to find out whether an appropriate mixture of these non-informative vectors could be identified but rather if, when sampling over a wide range of parameters, this effect that Cochran suggested does actually tend to produce better results using the F test rather than the Q test. Along these same lines Meyer (1967) has shown that the unconditional size of McNemar's test (Q test for J = 2) is less than nominal α unless "non-informative" vectors have probability zero. All of these considerations led this investigator to retain the ordinary degrees of freedom in the denominator of the F test used.

Results

In table 1 the reader will find summary results of the main portion of the investigation dealing with test size. The values in the table designated by AVE are averages of the proportions of times that the computed statistic exceeded the corresponding percentile of the appropriate reference distribution. The averages were taken over the 12 runs coming from the combinations of the four levels of interdependency among the observations in a block (ρ) and the three levels of success proportions (ϕ). The rows of this table denoted PCT are the average relative (percent) errors. Again each average is based on 12 relative errors and, for a given run, the relative error is $(p-\alpha)/\alpha$ where p is the empirical type I error and α is the nominal type I error. The quantity $\bar{E}$ is the average of these relative errors for these four selected upper percentiles.

Looking at the $\bar{E}$ values, an immediate observation is that, using this measure of fit of the empirical and theoretical distributions, the F test has a smaller average error than either Q or M for each design considered. It is also true that Q out performs M on this basis. Comparing Q and F for a moment we observe that the advantages of F over Q are largest for designs where the number of treatments is large relative to block size. For example, the largest difference in the $\bar{E}$ values is for the 10 treatment by 5 block design where $\bar{E}$ is 60% for the Q test and 29% for the F test.

Looking at the relative error as a function of nominal type I error we observe, as might be expected, that the errors increase as we go further out in the tail of the distribution, i.e., the fit of the empirical and theoretical distributions is poorer for α = .01 than .05. This trend is more noticeable for Q than for F and, to a large extent, this accounts for the smaller $\bar{E}$ for F for designs with two or three treatments. As a matter of fact, for two and three treatment designs, AVE and PCT values are almost identical for α = .10 and .05. The finding of Cochran that Q has

a positive bias above the 98th percentile is substantiated; the largest value of AVE for $\alpha = .01$ is .008 for the 10 x 20 design and most of the other values are substantially smaller. Although the corresponding value of AVE for the F test only achieves .01 for the 10 x 20 design, the majority of values for designs of modest size were as large as .008. Finally, the expected result that as the number of blocks increases the fit of the empirical and theoretical distributions is better, is observed

This writer somewhat arbitrarily selected the value of 20% relative error as a value which may be reasonably allowed in most educational experimentation. For tests at the 5% level this would correspond to <u>average</u> test size between .04 and .06. Considering 5% level F tests, the designs which satisfied this criterion led to the simple rule of those with 60 or more total observations, i.e., 2 x 30, 3 x 20, 3 x 30, 6 x 10, etc. As pointed out above, for $\alpha = .05$, F and Q do not differ for two and three treatments. It is also true that for six and ten treatments, the differences are minor so that, by bending the rule to allow PCT values of up to 22%, all of these designs satisfy the criterion for the Q test. As has been noted earlier, the F test tends to provide a better fit than Q in the upper tail of the distribution. However, for $\alpha = .01$, no simulations produced PCT values of less than 20% for either F or Q. For the F test, for the eight designs with 60 or more observations these relative errors range from 26% to 47% with a median of 35%. The corresponding minimum and maximum values and median for Q are 30%, 60% and 45%.

Not much has been said about the multivariate test using the test statistic denoted as M. It should be clear that, on the basis of the summary results presented in this table, the multivariate test has little to recommend it. As the discussion proceeds, certain characteristics of the multivariate test will be noted.

Although these results were somewhat encouraging, this writer noted that sample size recommendations based on these data would be misleading. The reason for this

is that, certain values of the parameters used in this investigation lead to "effective' sample sizes which are considerably less than the actual number of sample observations generated. The point is that "non-informative" responses, have no affect on the computation of either Q or the sums of squares for the F statistic. In Cochran's investigation of the small sample distribution of Q, he used eight different data sets, usually with about 3 or 4 treatments and 10 blocks. Cochran used only "informative" data in his investigation and found average errors of about 14% for .05 level test. Thus, with about 30 to 40 total observations, Cochran reported results which were somewhat better than those obtained here with 60 or more total observations.

In an attempt to bring the results of Cochran and those summarized here into closer agreement, this writer developed the notion of using "effective" sample size (N_e) as a criterion on which to categorize the 12 runs for each design. Effective sample size in this context refers to the number of "informative" response vectors generated. Since N_e is a random variable for a given set of parameters ρ and ϕ, the quantity which was selected to be used as a gross index of the number of "informative" responses was the expected or average value of N_e, which will be denoted $E(N_e)$.

The computation of $E(N_e)$ was carried out in the following manner. First, for a given configuration of ρ and ϕ, it is necessary to know the probability of a "non-informative" response (π_N). For some cases, this could have been done easily by hand. For example, for $\phi = .5$ and $\rho = .0$, the probability of either (0,0) or (1,1) response in a two treatments design is $.5^2 + .5^2 = .5$. This calculation is straightforward since $\rho = 0$ which, for the normal distribution, implies independence of the continuous latent variables. It is readily seen that the two binary variables are also independent. For those cases when ρ was not zero, the probability π_N of a "non-informative" response vector was estimated by generating 10,000 such samples on the computer. This method seemed to be sufficient considering the purposes for which this information was being obtained.

Using the value of Π_N, the effective sample size should follow a binomial distribution with the "nominal" sample size and $1-\Pi_N$ as parameters. For example for $\phi = .5$, $\rho = 0$ and "nominal" sample size of five, the distribution of N_e is:

Effective Sample Size (N_e)	0	1	2	3	4	5
Probability $Pr(N_e)$	1/32	5/32	10/32	10/32	5/32	1/32

The expected value of N_e was then computed in the usual fashion as $E(N_e) = \Sigma N_e \, Pr(N_e)$. In the example given this computation yields an expected sample size of 2.5. The careful reader will observe that, in some simulations, it is likely that all vectors will be non-informative. This situation leads to an indeterminate value for all three of the statistics. For the data on test size presented here, these data sets were taken as supporting the null hypothesis, and therefore, in the empirical size computations, 1000 is retained as the base.

By comparing the results of individual runs to the $E(N_e)$, this writer decided to isolate for further consideration those runs with $E(N_e)$ greater than six. That these runs are well behaved, is verified by Table 2 in which average empirical size results are given for those runs which satisfied the criterion. We note that the results are more in line with those presented by Cochran. In fact, the median relative error for Q for $\alpha = .05$ is 14%, the figure which Cochran reported. Again, the phenomena that F and Q have similar characteristics for $\alpha = .10$ is verified. For smaller values of α, however, the median average test size and relative errors for F and Q become increasingly different. For the two smallest values of nominal α, the relative error of Q is smaller than that of F in only one comparison.

For F the largest relative error for $\alpha = .025$ is 22% and most of the errors for $\alpha = .01$ are 30% or less. For the multivariate statistic we observe that for 3 treatments the procedure is fairly well behaved but that for 6 or more treatments, although AVE values are close to nominal α in some instances, percent errors are very large. This is due to a strange mixture of runs, most of which produce empirical type I error which either grossly exceed or underestimate the nominal values, but which produce

averages which are reasonably close to those values. One of the reasons that the results for the 6 and 10 treatment designs with 20 or fewer blocks lead to conservative procedures is that there were many instances when M was indeterminate. These were counted, the reader will recall, as instances for which the null hypothesis would be accepted.

An alternative method of determining whether there are any systematic tendencies for either of these procedures to be biased, is to count the instances in which the empirical proportion is larger than the nominal α. This was done for the Q and F test for each of the four values of α and the results are presented in Table 3. Since the multivariate test using M had exhibited such poor characteristics up to this point, results for it were not tabulated. An asterisk in this table signifies that, for more than half of the runs, empirical size exceeded nominal α. The columns headed "chance" in this table are simply one half the number of runs and indicate what would be expected if the nominal type I error values were the medians for the empirical proportions. The overall results indicate that, with the exception of the F test for $\alpha = .10$, for both procedures the empirical proportions tend to be less than the nominal values. However, the results for F are much nearer to the chance results for all nominal α. When these data are exhibited by number of treatment levels, the general statments made earlier are verified. For 2 or 3 treatments, the main advantage of F over Q is for α less than .05; for 6 or more treatments, F exhibits less bias over the whole range of α under consideration. The slight tendency for empirical size to exceed nominal α for F for 3 treatments at $\alpha = .10$ does not appear serious. There is also a tendency for the advantage of F over Q to diminish, except for small α, for designs with 30 observations. Two summary statements which seem in order are that (1) the F test provides a better approximation to nominal α for small α (α less than .05) and (2) for designs with the treatment to block ratio large (e.g., 6x10, 6x20, 10x10 and 10x20) the F distribution also provides a better fit for the other nominal type I errors under consideration.

Although results of individual runs will not be presented here because of the extensive number of tables which would be involved, a brief discussion of major points of information which they provide will be given. They are: (1) the runs excluded by the expected sample size greater than six criterion were the runs with large ρ and/or small ϕ values. For all three procedures described here, these tests were conservative. (2) For some smaller designs such as the 2x10, 2x20, 3x10 and 6x5, results using F will be reasonable if the data are not "pathological". That is, the fit of the variance ratio distribution appears to be adequate if the occurence of a success is not very rare and at the same time the variables are highly related. (3) Results for the M statistic are very inconsistent for different combinations of ρ and ϕ. For mildly correlated or uncorrelated data, the M statistic exhibited a very serious tendency for empirical size to grossly exceed nominal type I error. This tendency increased with block size and was not diminished as the number of blocks increased. Probably the most serious instance of this was for the 6x30 design and the $\rho = .0$, $\phi = .1$ run where the four empirical proportions were .226, 162, .101 and .049. The average relative error here is 273%. Although the writer has presented more complete information on the M statistic elsewhere (Mandeville, 1969), they have not been presented here in the interest of space and also because of the deficiencies already noted in the procedure.

Additional runs of test size were made for the three treatment and six treatment designs using non-patterned correlation matrices. These matrices were not chosen to be particularly exceptional, and, when taken in combination with the $\phi = .5$ values, yielded ε values which were approximately .97 for three treatments and .95 for six treatments. The quantity ε, introduced by Box (1954a, 1954b), is a measure of deviation from pattern and $\varepsilon = 1$ for patterned matrices. Sample sizes of 10 and 20 were used in combination with null success proportions of .1, .3 and .5 and these results are tabulated in tables4 and 5. Results of these runs are in reasonable

agreement with those for constant correlation matrices. For example, for three treatments the average correlation in the non-patterned matrix was near .5 so that comparisons with results for the runs with constant correlation of .5 are suggested. For the F statistic and the 3x20 design, the average errors are 24%, 18% and 11% for ϕ = .1, .3 and .5, respectively. Although not tabled here, the corresponding values for a constant α of .5 are 35%, 16% and 21%.

For the 6x10 design, however, the indication is that the non-patterned correlation structure lead to more conservative results for ϕ = .1 and .3 than the corresponding results for a patterned correlation matrix with ρ = .2. The average correlation in the six treatment correlation matrix is about .3.

Designs which were investigated with respect to empirical power included those studied as regards test size with the exception of designs with 5 and 30 blocks. Thus designs with 2, 3, 6 and 10 treatments were investigated for sample sizes of 10 and 20 blocks. Sample size 5 was eliminated since the results on test size were generally negative for such small samples unless a large number of treatments was involved. It was also the feeling of the writer that elimination of samples of size 30 would not greatly reduce the implications of this phase of the study.

Non-null vectors of proportions were obtained which exhibited linear departure about the central value ϕ_c = .5 and which would give theoretical (normal theory) power of .60 and .80 for tests run at the 5% and 1% levels. Although the normal theory assumptions were certainly not appropriate for the situation, this method was used so that, in the event that empirical power values were in agreement with the nominal values, it would be possible to recommend that a researcher use standard procedures for sample size computations. The constant correlation ρ was varied as before, taking the values of .0, .2, .5 and .8. To allow some generalization of the results, the combination ϕ_c = .3 and ρ = .2 was also included. For some sample sizes, no set of proportions between 0.0 and 1.0 could be found which satisfied certain size and power

combinations. This only occurred for ϕ_c = .3, however. In addition, the non-patterned correlation matrices were used in conjunction with ϕ_c = .5. Examples of some non-null proportions vectors used are given in Table 6. The reader interested in the details of the method used to obtain the non-null proportions vectors is referred to Mandeville (1969).

Because of the poor results obtained for the multivariate test, no power results will be presented here for the M statistic. Tables 7 through 10 contain power results for the Q and F tests. Dashes in these tables indicate that linear non-null proportions vectors do not exist for that combination of α, $1-\beta$ and the other parameters.

The results indicate that the F test is more powerful than Q for the designs with 2 or 3 treatments if a 1% level of significance is used and for either the 5% or 1% level for designs with 6 or more treatments. Of course, these results parallel those found in the earlier part of the investigation and are, therefore, not surprising. However, it is also noted that the F test yields empirical power which is in good agreement with, although generally slightly less than, normal theory power.

For the F statistic, the largest deviations of empirical from normal theory power results occurred for the two treatments designs. For both sample sizes, the largest average deviation occurred when normal theory power was .60 for 1% tests. These average empirical power values are .529 and .570 and represent deviations of .071 and .030 from the normal theory value of .60.

The ranges of the observed empirical power values were similar for F and Q and decreased for larger numbers of treatments so that they were seldom larger than about .050 for 6 and 10 treatments. This result is also consistent with the facts brought out earlier that Q is testing the more general hypothesis which includes treatment x block interaction and that the denominator of Q may be slightly inflated by the mean square for treatments. This effect should be most readily observed when the treatment mean square is large and the number of blocks is small, i.e., for α = .01 and I=10. Tables 7 - 10 verify that in these cases the advantages of F over Q is greatest.

The limited attempt to generalize the results using ϕ_c = .3 and the non-patterned correlation matrices have produced results which are in reasonable agreement with those for ϕ_c = .5 and patterned correlation matrices.

Summary and Conclusions

Considering test size, designs with 60 or more total observations were found to lead to average relative error for F and Q of about 20% or less for 5% tests. Results for F and Q were similar for α = .05 for 2 or 3 treatment designs. For designs with 6 or more treatments, the F test lead to to empirical size closer to nominal α than did the Q test. For α = .01 the F test out performed Q but relative errors for both statistics were often as large as 40%. When only designs and parameter specifications for which the average effective sample size was 6 or more were considered, the results were in good agreement with those reported by Cochran. For these cases the advantage of the F test for α = .01 was again observed. When non-patterned correlation matrices were used in conjunction with small null ϕ values, there was a slight tendency for the resulting test procedures to be more conservative than those with patterned correlation matrices.

As would be predicted from the results on test size, when power was considered, the F test proved to be more powerful than Q for α = .01 and for designs with 6 or more treatments this effect was observed for 5% tests also. The empirical power for F was generally slightly less than the nominal value but the results were close enough so that the use of standard parametric procedures to estimate sample size requirements seems justified.

This research was begun in hopes of allowing a recommendation that ANOVA procedure be used instead of Q for dichotomous data in a randomized block design. In addition to these two procedures a multivariate test was also considered. In the comparisons that have been made, the F test has:

1. Provided, for 5% tests, empirical size which has been as close to the nominal value, or closer to it than has been obtained for either Q or M.
2. Yielded empirical size closer to nominal size for $\alpha = .01$ than has been obtained for either Q or M.
3. Provided a maximum average percent error of about 20% for 5% tests when the total number of observations is 60 or more.
4. Yielded a median average relative error of about 10% for $\alpha = .05$ and 25% for $\alpha = .01$ for designs with average effective sample size of 6 or more.
5. Proved to be more powerful than Q for 1% tests for all designs considered.
6. Proved to be as powerful or more powerful than Q for 5% tests.
7. Yielded empirical power which was in good agreement with power predicted from normal theory calculations.

On the basis of these results the F test is recommended over Q or M when all of the following situations are met:

1. The researcher is mainly concerned with comparing average treatment effects.
2. Sixty or more total observations are available.
3. The interrelationships between the variables may be assumed to be reasonably constant.
4. The average success proportion is in the range .1 to .9.
5. The data might reasonably be thought of in terms of a normal ogive or logistic scaling model.
6. True type I errors may deviate by about 20% relative error for $\alpha = .05$ and by 40% or less for $\alpha = .01$ tests. By way of warning the reader should realize that, for either the F test or the Q test, certain large data sets can lead to results which deviate considerably from those obtained by the exact procedure. Thus, as pointed out by Tate and Brown, "When the true significanc level is needed, it would seem neccessary to construct the exact sampling

distribution." (1964, p. 18)

7. If power against linear non-null proportions vectors is of interest to the researcher, it is suggested that sample size computations based on normal theory considerations can be recommended.

The writer feels that these rules are somewhat conservative but suggest that further work possibly of an analytic nature, be done to determine the extent of the dependence of the results on points 3 and 5 above. It is hoped that work on generalizations to two treatment dimensions would also be forthcoming.

REFERENCES

Box, G. E. P. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems. 1) Effect of Inequalify of Variance in a One-Way Classification." Annals of Mathematical Statistics, 25, 1954a. 280-302

_____. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems. 2) Effects of Inequality of Variance and Covariance Between Errors in the Two-Way Classification." Annals of Mathematical Statistics, 25, 1954b. 484-498

Cochran, W. G. "The Comparison of Percentages in Matched Samples." Biometrics, 37, 1950. 256-266

D'Agostino, Ralph B. "A Second Look at Analysis of Variance on Dichotomous Data." Journal of Educational Measurement, 8, 1971. 327-332

Draper, John F. and Porter, Andrew C. "A Test for a Neglected Source of Variation: the Individual Differences by Repeated Measures Interaction." Paper presented at the annual meeting of American Educational Research Association, Minneapolis Minnesota, 1970.

Fleiss, J. L. "Estimating the Accuracy of Dichotomous Judgments." Psychometrika, 30, 1965. 256-266

Hsu, Tse-Chi and Feldt, Leonard S. "The Effect of Limitations on the Number of Criterion Score Values on the Significance Level of the F-Test." American Educational Research Journal, 6, 1969. 515-527

Lunney, G. H. "Using Analysis of Variance With a Dichotomous Dependent Variable: an Empirical Study." Journal of Educational Measurement, 4, 1970. 263-269

Mandeville, Garrett K. "A Monte Carlo Investigation of the Adequacy of Standard Analysis of Variance Test Procedures for Dependent Binary Variates." Unpublished Ph.D. Thesis, University of Minnesota, 1969.

McNemar, Q. Psychological Statistics. New York, John Wiley and Sons, 1962.

Meyer, Donald. "The Unconditional Power Function of McNemar's Test." Paper presented at annual meeting of the American Educational Research Association, New York, 1967.

Porter, Andrew C. and McSweeney, Maryellen. "Randomized Blocks Design and Non-parametric Statistics." Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, 1970.

Rao, C. R. Linear Statistical Inference and Its Applications, New York, John Wiley and Sons, 1965.

Seegar, Paul and Gabrielsson, Alf. "Applicability of the Cochran Q Test and the F Test for Statistical Analysis of Dichotomous Data for Dependent Samples." Psychological Bulletin, 69, 1968. 269-277

Siegel, Sidney. Non-Parametric Statistics for the Behavior Sciences. New York, McGraw-Hill, 1956.

Tate, M. W. and Brown, Sara. Tables for Comparing Related-Sample Percentages, University of Pennsylvania Graduate School of Education, 1964.

_____. "Note on the Cochran Q Test." Journal of the American Statistical Association, 65, 1970. 155-160

Table 1 -- Average empirical test size and relative percent error for all designs. Values have been averaged over the 12 runs for each design.

Statistic		Q					F					M				
Design		Nom. type I error .10	.05	.025	.01	$\bar{E}$	Nom. type I error .10	.05	.025	.01	$\bar{E}$	Nom. type I error .10	.05	.025	.01	$\bar{E}$
2x5	AVE	035	005	000	000	89	035	005	005	000	84	Identical to F				
	PCT	65	90	100	100		65	90	81	100						
2x10	AVE	075	025	004	001	65	075	025	010	004	57					
	PCT	32	52	85	92		32	52	68	75						
2x20	AVE	099	038	011	003	42	099	039	017	004	35					
	PCT	10	24	62	73		10	24	40	65						
2x30	AVE	101	047	016	005	28	101	047	022	008	23					
	PCT	8	16	40	48		8	16	19	47						
3x5	AVE	053	018	002	000	76	059	018	007	002	68	025	005	000	000	91
	PCT	50	64	93	98		51	65	73	83		75	90	100	100	
3x10	AVE	074	032	009	002	52	077	032	011	004	45	051	020	007	003	65
	PCT	28	36	65	79		28	36	57	59		51	65	71	72	
3x20	AVE	098	047	018	005	25	101	047	021	008	23	092	039	018	006	43
	PCT	9	19	28	45		9	19	28	38		24	38	44	67	
3x30	AVE	097	046	017	006	23	098	046	019	008	18	099	044	021	006	32
	PCT	5	11	31	43		5	11	24	33		14	26	36	52	

Table 1 -- (continued)

6x5	AVE	059	018	005	001	69	076	029	015	006	44	Not defined				
	PCT	42	65	80	90		32	47	46	52						
6x10	AVE	078	034	012	004	42	086	042	020	008	23	032	016	007	003	77
	PCT	22	33	51	60		15	20	25	32		68	75	78	89	
6x20	AVE	095	043	019	006	25	100	046	023	009	21	081	039	020	007	62
	PCT	7	18	29	46		7	15	24	38		46	57	70	76	
6x30	AVE	093	046	021	007	18	096	049	023	008	13	100	055	027	010	61
	PCT	8	11	18	37		7	7	14	26		34	54	70	86	
10x5	AVE	061	023	008	002	60	078	038	021	008	29	Not defined				
	PCT	39	53	68	78		24	27	30	35						
10x10	AVE	081	036	016	005	37	090	045	022	009	27	003	000	000	000	99
	PCT	21	28	42	56		16	19	30	42		97	99	100	100	
10x20	AVE	090	043	021	008	19	094	047	024	010	15	057	026	014	006	68
	PCT	13	15	19	30		11	11	10	29		59	64	70	80	

Table 2 -- Average empirical test size and relative percent error for designs and combinations of parameters which lead to expected sample sizes of 6 or more.

Design	Number of runs		O Nom. Type I Error .10	.05	.025	.01	E	F Nom. Type I Error .10	.05	.025	.01	E	M Nom. Type I Error .10	.05	.025	.01	E
2x20	5	AVE	107	045	021	006	21	107	047	026	009	14					
		PCT	8	12	25	40		8	11	17	20		Identical to F				
2x30	7	AVE	100	047	021	007	19	100	047	023	011	18					
		PCT	8	15	22	29		8	15	16	31						
3x10	3	AVE	092	046	017	005	26	101	046	024	009	10	098	050	021	010	14
		PCT	8	9	32	53		4	9	13	13		8	20	16	10	
3x20	8	AVE	105	051	022	007	17	107	051	026	010	16	100	046	021	009	34
		PCT	6	16	16	29		7	16	16	25		15	30	34	55	
3x30	10	AVE	098	046	018	007	20	099	046	021	009	16	105	048	023	007	26
		PCT	5	11	26	38		5	11	18	29		11	20	30	43	
6x10	6	AVE	089	043	016	005	27	099	051	025	010	10	061	032	014	006	56
		PCT	11	16	36	47		3	8	11	20		39	51	57	78	
6x20	11	AVE	097	044	020	007	22	102	048	024	009	19	088	042	021	007	59
		PCT	4	16	27	42		6	13	22	36		41	53	67	74	
6x30	11	AVE	094	046	022	007	17	096	049	024	009	12	109	060	029	011	58
		PCT	8	10	16	33		8	6	12	23		29	56	67	85	
10x10	7	AVE	094	043	019	006	24	101	053	028	011	18	005	001	000	000	99
		PCT	13	16	29	39		8	9	22	30		95	99	100	100	
10x20	11	AVE	092	043	021	008	17	095	048	025	011	13	062	028	015	007	65
		PCT	12	13	16	25		10	10	8	24		55	61	67	78	
Median		AVE	0955	0455	0205	007	20.5	1005	048	0245	0095	15	099	0465	022	008	45
		PCT	8	14	25.5	38.5		7.5	10.5	16	24.5		22	41	45	65	

For the 2x5, 2x10, 3x5, 6x5 and 10x5 designs, none of the 12 runs had an expected sample size of 6 or greater.

Table 3 -- Frequency of occurrence of empirical test size greater than nominal α for runs with expected sample size of six or more.

		Nominal Type I Error														
	Number	.10			.05			.025			.01			Total		
Design	of runs	Q	F	Chance	Q	F	Chance	Q	F	Chance	Q	F	Chance	Q	F	Chance
2x20	5	4*	4*	2.5	1	2	2.5	1	3*	2.5	0	1	2.5	6	10	10
2x30	7	3	3	3.5	2	2	3.5	1	2	3.5	0	5*	3.5	6	12	11
3x10	3	0	2*	1.5	0	0	1.5	0	1	1.5	0	1	1.5	0	4	6
3x20	8	4	8*	4	4	4	4	2	4	4	0	3	4	10	19*	16
3x30	10	4	4	5	3	3	5	0	1	5	1	3	5	8	11	20
6x10	6	0	2	3	1	3	3	0	3	3	0	2	3	1	10	12
6x20	11	4	8*	5.5	2	5	5.5	2	5	5.5	1	4	5.5	9	22	22
6x30	11	2	3	5.5	2	3	5.5	2	3	5.5	3	3	5.5	9	12	22
10x10	7	1	3	3.5	1	4*	3.5	2	3	3.5	1	3	3.5	5	13	14
10x20	11	3	5	5.5	0	4	5.5	1	6*	5.5	2	5	5.5	6	20	22
TOTAL	79	25	42*	39.5	16	30	39.5	11	31	39.5	8	30	39.5	60	133	158

An asterisk(*) in the above table signifies that, for more than half of the runs, empirical test size exceeded nominal α.

Table 4 -- Empirical upper tail probabilities for three test statistics for the 3x10 and 3x20 designs using a non-patterned correlation matrix.

Sample Size	Statistic	ϕ	Nominal Type I Error .10	.05	.025	.01	E
		.1	044	013	001	000	82
	Q	.3	099	C53	012	004	30
		.5	111	052	016	004	28
		.1	044	013	001	000	82
10	F	.3	101	053	014	005	23
		.5	116	052	021	007	17
		.1	004	001	000	000	99
	M	.3	055	019	008	002	64
		.5	065	018	003	001	69
		.1	090	043	019	003	30
	Q	.3	094	051	026	008	8
		.5	104	058	022	004	23
		.1	091	043	019	005	24
20	F	.3	098	051	029	015	18
		.5	105	058	028	011	11
		.1	053	013	004	000	76
	M	.3	098	039	017	007	22
		.5	117	051	025	009	7

Table 5 -- Empirical upper tail probabilities for three test statistics for the 6x10 and 6x20 designs using a non-patterned correlation matrix.

Sample Size	Statistic	ϕ	Nominal Type I Error .10	.05	.025	.01	E
10	Q	.1	069	021	007	001	63
		.3	098	051	026	011	5
		.5	073	034	016	005	36
	F	.1	073	031	013	004	41
		.3	107	065	031	015	28
		.5	083	040	025	008	14
	M	.1	000	000	000	000	100
		.3	045	017	006	003	67
		.5	047	015	007	000	74
20	Q	.1	068	036	012	004	43
		.3	098	051	023	009	6
		.5	089	045	018	010	12
	F	.1	073	039	015	006	32
		.3	103	055	031	013	17
		.5	093	048	022	012	11
	M	.1	053	017	003	000	75
		.3	102	048	021	009	8
		.5	092	039	017	008	21

Table 6 -- Examples of non-null proportions vectors used in the investigation of empirical power.

Design	ϕ_c	ρ	α	$1-\beta$	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
2 x 10	.5	.2	.05	.60	.268	.732				
	.5	.2	.05	.80	.222	.778				
3 x 10	.5	.2	.05	.60	.240	.500	.760			
	.5	.2	.01	.60	.178	.500	.822			
	.3	.2	.05	.60	.061	.300	.539			
	.5	NP*	.05	.60	.262	.500	.737			
3 x 20	.5	.2	.05	.60	.317	.500	.683			
6 x 10	.5	.2	.05	.60	.242	.345	.448	.552	.655	.758
	.3	.2	.05	.60	.061	.157	.252	.348	.443	.539
6 x 20	.5	.2	.05	.60	.317	.390	.464	.536	.610	.683

*In this and succeeding tables where NP appears it refers to the appropriate non-patterned correlation matrix.

Table 7 -- Empirical power for the Q and F tests for designs with 2 treatments. Non-null proportions vectors would yield theory power for the F test of .60 and .80 at each of α = .05 and .01.

Statistic			Q				F			
Nominal α			.05		.01		.05		.01	
Sample Size	ϕ_c	ρ	Nominal Power for F .60	.80	.60	.80	Nominal Power for F .60	.80	.60	.80
10	.5	.0	559	688	421	583	559	688	586	754
	.5	.2	552	761	352	547	552	761	553	753
	.5	.5	600	785	290	508	600	785	525	732
	.5	.8	565	826	258	481	565	826	476	708
	.3	.2	598	796	279	-	598	796	507	-
	AVE		575	771	320	530	575	771	529	737
	RANGE		048	138	163	102	041	138	110	046
20	.5	.0	523	751	509	691	561	774	612	771
	.5	.2	560	746	494	699	574	768	596	786
	.5	.5	553	784	474	730	554	785	547	797
	.5	.8	625	846	490	754	625	846	508	763
	.3	.2	586	778	518	736	594	781	587	793
	AVE		569	781	497	722	582	791	570	782
	RANGE		102	100	044	063	071	078	104	034

Note: On this and succeeding tables the dash "-" indicates that, due to the restrictions on the ϕ-values, no non-null vector exists.

Table 8 -- Empirical power for the Q and F tests for design with 3 treatments. Non-null proportions vectors would yield normal theory power for the F test of .60 and .80 at each of α = .05 and .01.

Statistic			Q				F			
Nominal α			.05		.01		.05		.01	
Sample Size	ϕ_c	ρ	Nominal Power for F .60	.80	.60	.80	Nominal Power for F .60	.80	.60	.80
10	.5	.0	586	778	510	685	586	778	599	762
	.5	.2	599	762	473	672	599	762	568	745
	.5	.5	586	823	455	657	586	823	556	763
	.5	.8	576	811	397	660	576	811	545	772
	.3	.2	616	810	-	-	616	810	-	-
	.5	NP	586	799	462	684	586	799	576	769
	AVE		591	797	459	672	591	797	569	762
	RANGE		040	061	113	028	030	061	054	027
20	.5	.0	614	800	568	732	614	800	613	779
	.5	.2	619	775	549	772	619	775	603	809
	.5	.5	591	803	525	749	591	803	586	798
	.5	.8	632	806	561	768	632	806	623	804
	.3	.2	586	810	554	754	586	810	609	802
	.5	NP	588	791	540	728	588	791	598	768
	AVE		605	798	550	751	605	798	605	793
	RANGE		046	035	043	044	046	035	037	041

Table 9 -- Empirical power for the Q and F tests for designs with 6 treatments. Non-null proportions vectors would yield normal theory power for the F test of .60 and .80 at each of α = .05 and .01.

Statistic			Q				F			
Nominal α			.05		.01		.05		.01	
Sample Size	ϕ_c	ρ	Nominal Power for F .60	.80	.60	.80	Nominal Power for F .60	.80	.60	.80
10	.5	.0	580	747	500	729	610	786	584	818
	.5	.2	552	765	502	725	594	799	599	805
	.5	.5	543	769	514	721	590	791	612	791
	.5	.8	496	732	471	683	536	765	572	768
	.3	.2	568	778	473	-	607	806	581	-
	.5	NP	557	758	531	741	599	770	598	792
	AVE		549	758	499	720	589	786	591	795
	RANGE		084	046	060	058	074	041	040	050
20	.5	.0	604	801	559	755	619	809	597	788
	.5	.2	573	777	563	746	587	794	612	773
	.5	.5	577	769	584	750	592	789	624	780
	.5	.8	574	797	566	780	601	810	610	811
	.3	.2	615	793	541	771	632	808	597	801
	.5	NP	588	786	522	747	609	794	556	775
	AVE		589	787	556	757	607	801	599	788
	RANGE		042	032	062	044	045	021	068	038

Table 10 -- Empirical power for the Q and F tests for designs with 10 treatments. Non-null proportions vectors would yield normal theory power for the F test of .60 and .80 at each of α = .05 and .01.

Statistic			Q				F			
Nominal α			.05		.01		.05		.01	
Sample Size	ϕ_c	ρ	Nominal Power for F .60	.80	.60	.80	Nominal Power for F .60	.80	.60	.80
10	.5	.0	559	758	533	720	587	788	597	774
	.5	.2	542	767	525	734	579	794	597	795
	.5	.5	568	764	532	715	611	799	602	799
	.5	.8	515	723	482	707	548	757	578	749
	.3	.2	540	756	514	-	580	790	617	-
	AVE		545	754	517	719	581	786	598	779
	RANGE		053	044	050	027	063	042	039	050
20	.5	.0	576	786	557	771	593	800	588	796
	.5	.2	567	780	559	768	589	794	590	797
	.5	.5	590	756	563	799	609	765	590	816
	.5	.8	578	771	584	761	592	781	611	784
	.3	.2	594	786	548	767	613	794	585	793
	AVE		581	776	562	773	599	787	593	797
	RANGE		027	030	036	038	024	035	026	032